Journal of Mathematical Psychology 72 (2016) 19–32

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology

Alexander Ly *, Josine Verhagen, Eric-Jan Wagenmakers

University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, The Netherlands

HIGHLIGHTS

- The Bayes factor follows logically from Jeffreys's philosophy of model selection.
- The ideas are illustrated with two examples: the Bayesian t-test and correlation test.
- The Bayes factors are adapted to one-sided tests.
- The Bayes factors are illustrated with various applications in psychological research.

ARTICLE INFO

Article history:
Available online 28 August 2015

Keywords:
Model selection
Bayes factors
Harold Jeffreys

ABSTRACT

Harold Jeffreys pioneered the development of default Bayes factor hypothesis tests for standard statistical problems. Using Jeffreys's Bayes factor hypothesis tests, researchers can grade the decisiveness of the evidence that the data provide for a point null hypothesis $\mathcal{H}_0$ versus a composite alternative hypothesis $\mathcal{H}_1$. Consequently, Jeffreys's tests are of considerable theoretical and practical relevance for empirical researchers in general and for experimental psychologists in particular. To highlight this relevance and to facilitate the interpretation and use of Jeffreys's Bayes factor tests we focus on two common inferential scenarios: testing the nullity of a normal mean (i.e., the Bayesian equivalent of the t-test) and testing the nullity of a correlation. For both Bayes factor tests, we explain their development, we extend them to one-sided problems, and we apply them to concrete examples from experimental psychology.

Consider the common scenario where a researcher entertains two competing hypotheses. One, the null hypothesis $\mathcal{H}_0$, is implemented as a statistical model that stipulates the nullity of a parameter of interest (i.e., $\mu = 0$); the other, the alternative hypothesis $\mathcal{H}_1$, is implemented as a statistical model that allows the parameter of interest to differ from zero. How should one quantify the relative support that the observed data provide for $\mathcal{H}_0$ versus $\mathcal{H}_1$? Harold Jeffreys argued that this is done by assigning prior mass to the point null hypothesis (or "general law") $\mathcal{H}_0$, and then calculate the degree to which the data shift one's prior beliefs about the relative plausibility of $\mathcal{H}_0$ versus $\mathcal{H}_1$. The factor by which the data shift one's prior beliefs about the relative plausibility of two competing models is now widely known as the Bayes factor, and it is arguably the gold standard for Bayesian model comparison and hypothesis testing (e.g., Berger, 2006; Lee & Wagenmakers, 2013; Lewis & Raftery, 1997; Myung & Pitt, 1997; O'Hagan & Forster, 2004).

In his brilliant monograph "Theory of Probability", Jeffreys introduced a series of default Bayes factor tests for common statistical scenarios. Despite their considerable theoretical and practical appeal, however, these tests are hardly ever used in experimental psychology and other empirical disciplines. A notable exception concerns Jeffreys's equivalent of the t-test, which has recently been promoted by Jeffrey Rouder, Richard Morey, and colleagues (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009). One of the reasons for the relative obscurity of Jeffreys's default tests may be that a thorough understanding of "Theory of Probability" requires not only an affinity with mathematics but also a willingness to decipher Jeffreys's non-standard notation.

In an attempt to make Jeffreys's default Bayes factor tests accessible to a wider audience we explain the basic principles that drove their development and then focus on two popular inferential scenarios: testing the nullity of a normal mean (i.e., the Bayesian t-test) and testing the nullity of a correlation. We illustrate Jeffreys's methodology using data sets from psychological studies. The paper is organized as follows: The first section provides some historical background and outlines four of Jeffreys's convictions regarding scientific learning. The second section shows how the

* Corresponding author.
E-mail address: alexanderly.nl@gmail.com (A. Ly).

http://dx.doi.org/10.1016/j.jmp.2015.06.004
0022-2496/

Bayes factor is a natural consequence of these four convictions. We decided to include Jeffreys's own words where appropriate, so as to give the reader an accurate impression of Jeffreys's ideas as well as his compelling style of writing. The third section presents the procedure from which so-called default Bayes factors can be constructed. This procedure is illustrated with the redevelopment of the Bayesian counterpart for the t-test and the Bayesian correlation test. For both the t-test and the correlation test, we also derive one-sided versions of Jeffreys's original tests. We apply the resulting Bayes factors to data sets from psychological studies. The last section concludes with a summary and a discussion.

1. Historical and philosophical background of the Bayes factor

1.1. Life and work

Sir Harold Jeffreys was born in 1891 in County Durham, United Kingdom, and died in 1989 in Cambridge. Jeffreys first earned broad academic recognition in geophysics when he discovered the earth's internal structure (Bolt, 1982; Jeffreys, 1924). In 1946, Jeffreys was awarded the Plumian Chair of Astronomy, a position he held until 1958. After his "retirement" Jeffreys continued his research to complete a record-breaking 75 years of continuous academic service at any Oxbridge college, during which he was awarded medals by the geological, astronomical, meteorological, and statistical communities (Cook, 1990; Huzurbazar, 1991; Lindley, 1991; Swirles, 1991). His mathematical ability is on display in the book "Methods of Mathematical Physics", which he wrote together with his wife (Jeffreys & Jeffreys, 1946).

Our first focus is on the general philosophical framework for induction and statistical inference put forward by Jeffreys in his monographs "Scientific Inference" (Jeffreys, 1931, second edition 1955, third edition 1973) and "Theory of Probability" (henceforth ToP; first edition 1939, second edition 1948, third edition 1961). An extended modern summary of ToP is provided by (Robert, Chopin, & Rousseau, 2009). Jeffreys's ToP rests on a principled philosophy of scientific learning (ToP, Chapter I). In ToP, Jeffreys distinguishes sharply between problems of parameter estimation and problems of hypothesis testing. For estimation problems, Jeffreys outlines his famous transformation-invariant "Jeffreys's priors" (ToP, Chapter III); for testing problems, Jeffreys proposes a series of default Bayes factor tests to grade the support that observed data provide for a point null hypothesis $\mathcal{H}_0$ versus a composite $\mathcal{H}_1$ (ToP, Chapter V). A detailed summary of Jeffreys's contributions to statistics is available online at www.economics.soton.ac.uk/staff/aldrich/jeffreysweb.htm.

For several decades, Jeffreys was one of only few scientists who actively developed, used, and promoted Bayesian methods. In recognition of Jeffreys's persistence in the face of relative isolation, E. T. Jaynes's dedication of his own book, "Probability theory: The logic of science", reads: "Dedicated to the memory of Sir Harold Jeffreys, who saw the truth and preserved it" (Jaynes, 2003). In 1980, the seminal work of Jeffreys was celebrated in the 29-chapter book "Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys" (e.g., Geisser, 1980; Good, 1980; Lindley, 1980; Zellner, 1980). In one of its chapters, Dennis Lindley discusses ToP and argues that "The *Theory* is a wonderfully rich book. Open it at almost any page, read carefully, and you will discover some pearl" (Lindley, 1980, p. 37).

Despite discovering the internal structure of the earth and proposing a famous rule for developing transformation-invariant prior distributions, Jeffreys himself considered his greatest scientific achievement to be the development of the Bayesian hypothesis test by means of default Bayes factors (Senn, 2009). In what follows, we explain the rationale behind Jeffreys's Bayes factors and demonstrate their use for two concrete tests.

1.2. Jeffreys's view of scientific learning

Jeffreys developed his Bayes factor hypothesis tests as a natural consequence of his perspective on statistical inference, a philosophy guided by principles and convictions inspired by Karl Pearson's classic book *The Grammar of Science* and by the work of W. E. Johnson and Dorothy Wrinch. Without any claim to completeness or objectivity, here we outline four of Jeffreys's principles and convictions that we find particularly informative and relevant.

1.2.1. Conviction i: Inference is inductive

Jeffreys's first conviction was that scientific progress depends primarily on induction (i.e., learning from experience). For instance, he states "There is a solid mass of belief reached inductively, ranging from common experience and the meanings of words, to some of the most advanced laws of physics, on which there is general agreement among people that have studied the data" (Jeffreys, 1955, p. 276) and, similarly: "When I taste the contents of a jar labelled 'raspberry jam' I expect a definite sensation, inferred from previous instances. When a musical composer scores a bar he expects a definite set of sounds to follow when an orchestra plays it. Such inferences are not deductive, nor indeed are they made with certainty at all, though they are still widely supposed to be" (Jeffreys, 1973, p. 1). The same sentiment is stated more forcefully in ToP: "(...) the fact that deductive logic provides no explanation of the choice of the simplest law is an absolute proof that deductive logic is grossly inadequate to cover scientific and practical requirements" (Jeffreys, 1961, p. 5). Hence, inference is inductive and should be guided by the data we observe.

1.2.2. Conviction ii: Induction requires a logic of partial belief

Jeffreys's second conviction is that in order to formalize induction one requires a logic of partial belief: "The idea of a reasonable degree of belief intermediate between proof and disproof is fundamental. It is an extension of ordinary logic, which deals only with the extreme cases" (Jeffreys, 1955, p. 275). This logic of partial belief, Jeffreys showed, needs to obey the rules of probability calculus in order to fulfill general desiderata of consistent reasoning—thus, degrees of belief can be thought of as probabilities (cf. Ramsey, 1926). Hence, all the unknowns should be instantiated as random variables by specifying so-called prior distributions before any datum is collected. Using Bayes' theorem, these priors can then be updated to posteriors conditioned on the data that were actually observed.

1.2.3. Conviction iii: The test of a general law requires it be given prior probability

Jeffreys's third conviction stems from his rejection of treating a testing issue as one of estimation. This is explained clearly and concisely by Jeffreys himself:

> "My chief interest is in significance tests. This goes back to a remark in Pearson's *Grammar of Science* and to a paper of 1918 by C. D. Broad. Broad used Laplace's theory of sampling, which supposes that if we have a population of n members, r of which may have a property ϕ, and we do not know r, the prior probability of any particular value of r (0 to n) is $1/(n+1)$. Broad showed that on this assessment, if we take a sample of number m and find them all with ϕ, the posterior probability that all n are ϕs is $(m+1)/(n+1)$. A general rule would never acquire a high probability until nearly the whole of the class had been inspected. We could never be reasonably sure that apple trees would always bear apples (if anything). The result is preposterous, and started the work of Wrinch and myself in 1919–1923. Our point was that giving prior probability

> $1/(n+1)$ to a general law is that for n large we are already expressing strong confidence that no general law is true. The way out is obvious. To make it possible to get a high probability for a general law from a finite sample the prior probability must have at least some positive value independent of n" (Jeffreys, 1980, p. 452).

The allocation of probability to the null hypothesis is known as the simplicity postulate (Wrinch & Jeffreys, 1921), that is, the notion that scientific hypotheses can be assigned prior plausibility in accordance with their complexity, such that "the simpler laws have the greater prior probabilities" (e.g., Jeffreys, 1961, p. 47; see also Jeffreys, 1973, p. 38). In the case of testing a point null hypothesis, the simplicity postulate expresses itself through the recognition that the point null hypothesis represents a general law and, hence, requires a separate, non-zero prior probability.

Jeffreys's view of the null hypothesis as a general law is influenced by his background in (geo)physics. For instance, Newton's law of gravity postulates the existence of a fixed universal gravitational constant G. Clearly, this law is more than just a statement about a constant; it provides a model of motion that relates data to parameters. In this context, the null hypothesis should be identified with its own separate null model $\mathcal{M}_0$ rather than be perceived as a simplified statement $\mathcal{H}_0$ within the model $\mathcal{M}_1$.

Hence, Jeffreys's third conviction holds that in order to test the adequacy of a null hypothesis, the model that instantiates that hypothesis needs to be assigned a separate prior probability, which can be updated by the data to a posterior probability.

1.2.4. Conviction iv: Classical tests are inadequate

Jeffreys's fourth conviction was that classical "Fisherian" p-values are inadequate for the purpose of hypothesis testing. In the preface to the first edition of ToP, Jeffreys outlines the core problem: "Modern statisticians have developed extensive mathematical techniques, but for the most part have rejected the notion of the probability of a hypothesis, and thereby deprived themselves of any way of saying precisely what they mean when they decide between hypotheses" (Jeffreys, 1961, p. ix). Specifically, Jeffreys pointed out that the p-value significance test "(...) does not give the probability of the hypothesis; what it does give is a convenient, though rough, criterion of whether closer investigation is needed" (Jeffreys, 1973, p. 49). Thus, by selectively focusing on the adequacy of predictions under the null hypothesis – and by neglecting the adequacy of predictions under the alternative hypotheses – researchers may reach conclusions that are premature (see also the Gosset–Berkson critique, Berkson, 1938; Wagenmakers, Verhagen, Ly, Matzke, Steingroever, Rouder and Morey, 2015):

> "Is it of the slightest use to reject a hypothesis until we have some idea of what to put in its place? If there is no clearly stated alternative, and the null hypothesis is rejected, we are simply left without any rule at all, whereas the null hypothesis, though not satisfactory, may at any rate show some sort of correspondence with the facts" (Jeffreys, 1961, p. 390).

Jeffreys also argued against the logical validity of p-values, famously pointing out that they depend on more extreme events that have not been observed: "What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure" (Jeffreys, 1961, p. 385). In a later paper, Jeffreys clarifies this statement: "I have always considered the arguments for the use of P absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened" (Jeffreys, 1980, p. 453).

In sum, Jeffreys was convinced that induction is an extended form of logic; that this "logic of partial beliefs" needs to treat degrees of belief as probabilities; that simple laws or hypotheses should be viewed as separate models that are allocated non-zero prior probabilities, and that a useful and logically consistent method of hypothesis testing need to be comparative, and needs to be based on the data at hand rather than on data that were never observed. These convictions coalesced in Jeffreys's development of the Bayes factor, an attempt to provide a consistent method of model selection and hypothesis testing that remedies the weaknesses and limitations inherent to p-value statistical hypothesis testing.

1.3. The Bayes factor hypothesis test

In reverse order, we elaborate on the way in which each of Jeffreys's convictions motivated the construction of his Bayes factor alternative to the classical hypothesis test.

1.3.1. ad. Conviction iv: Classical tests are inadequate

Jeffreys's development of a Bayesian hypothesis test was motivated in part by his conviction that the use of classical p values is "absurd". Nevertheless, Jeffreys reported that the use of Bayes factors generally yields conclusions similar to those reached by means of p values: "As a matter of fact I have applied my significance tests to numerous applications that have also been worked out by Fisher's, and have not yet found a disagreement in the actual decisions reached" (Jeffreys, 1961, p. 393); thus, "In spite of the difference in principle between my tests and those based on the P integrals (...) it appears that there is not much difference in the practical recommendations" (Jeffreys, 1961). However, Jeffreys was acutely aware of the fact that disagreements can occur (see also Edwards, Lindman, & Savage, 1963; Lindley, 1957). In psychology, these disagreements appear to arise repeatedly, especially for cases in which the p value is in the interval from .01 to .05 (Johnson, 2013; Wetzels et al., 2011).

1.3.2. ad. Conviction iii: The test of a general law requires it be given prior probability

Jeffreys first identified the null hypothesis with a separate null model $\mathcal{M}_0$ that represents a general law and pits it against the alternative model $\mathcal{M}_1$ which relaxes the restriction imposed by the law. For instance, for the t-test, $\mathcal{M}_0$: normal data X with $\mu = 0$ – the law says that the population mean is zero – and $\mathcal{M}_1$: normal data X that allows μ to vary freely. As we do not know whether the data were generated according to $\mathcal{M}_0$ or $\mathcal{M}_1$ we consider the model choice a random variable such that $P(\mathcal{M}_1) + P(\mathcal{M}_0) = 1$.

1.3.3. ad. Conviction ii: Induction requires a logic of partial belief

As the unknowns are considered to be random, we can apply Bayes' rule to yield posterior model probabilities given the observed data, as follows:

$$P(\mathcal{M}_1 \mid d) = \frac{p(d \mid \mathcal{M}_1)P(\mathcal{M}_1)}{P(d)}, \tag{1}$$

$$P(\mathcal{M}_0 \mid d) = \frac{p(d \mid \mathcal{M}_0)P(\mathcal{M}_0)}{P(d)}, \tag{2}$$

where $p(d \mid \mathcal{M}_i)$ is known as the marginal likelihood which represents the "likelihood of the data being generated from model $\mathcal{M}_i$". By taking the ratio of the two expressions above, the common term $P(d)$ drops out yielding the key expression:

$$\underbrace{\frac{P(\mathcal{M}_1 \mid d)}{P(\mathcal{M}_0 \mid d)}}_{\text{Posterior odds}} = \underbrace{\frac{p(d \mid \mathcal{M}_1)}{p(d \mid \mathcal{M}_0)}}_{\text{BF}_{10}(d)} \underbrace{\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_0)}}_{\text{Prior odds}}. \tag{3}$$

This equation has three crucial ingredients. First, the prior odds quantifies the relative plausibility of $\mathcal{M}_1$ over $\mathcal{M}_0$ before any datum is observed. Most researchers enter experiments with prior knowledge, prior experiences, and prior expectations, and these can in principle be used to determine the prior odds. Jeffreys preferred the assumption that both models are equally likely a priori, such that $P(\mathcal{M}_0) = P(\mathcal{M}_1) = 1/2$. This is consistent with the Wrinch–Jeffreys simplicity postulate in the sense that prior mass $1/2$ is assigned to a parsimonious model (e.g., $\mathcal{M}_0 : \mu = 0$, the general law), and the remaining $1/2$ is spread out over a larger model $\mathcal{M}_1$ where μ is unrestricted. In general then, the prior odds quantify a researcher's initial skepticism about the hypotheses under test. The second ingredient is the posterior odds, which quantifies the relative plausibility of $\mathcal{M}_0$ and $\mathcal{M}_1$ after having observed data d. The third ingredient is the Bayes factor (Jeffreys, 1935): the extent to which data d update the prior odds to the posterior odds. For instance, when $\text{BF}_{10}(d) = 9$, the observed data d are 9 times more likely to have occurred under $\mathcal{M}_1$ than under $\mathcal{M}_0$; when $\text{BF}_{10}(d) = 0.2$, the observed data d are 5 times more likely to have occurred under $\mathcal{M}_0$ than under $\mathcal{M}_1$. The Bayes factor, thus, quantifies the relative probability of the observed data under each of the two competing hypotheses.

Typically, each model $\mathcal{M}_i$ has unknown parameters θ_i that, in accordance to Jeffreys's second conviction, are considered as random with a density given by $\pi_i(\theta_i)$. By the law of total probability the "likelihood of the data being generated from model $\mathcal{M}_i$" is then calculated by integrating out the unknown parameters within that model, that is, $p(d \mid \mathcal{M}_i) = \int f(d \mid \theta_i, \mathcal{M}_i)\pi_i(\theta_i)\mathrm{d}\theta_i$, where $f(d \mid \theta_i, \mathcal{M}_i)$ is the likelihood, that is, the function that relates the observed data to the unknown parameters θ_i within model $\mathcal{M}_i$ (e.g., Myung, 2003). Hence, when we do not know which of two models $(\mathcal{M}_0, \mathcal{M}_1)$ generated the observed data and both models contain unknown parameters, we have to specify two prior densities (π_0, π_1) which formalize our uncertainty before any datum has been observed.

1.3.4. ad. Conviction i: Inference is inductive

The specification of the two prior distributions π_0, π_1 is guided by two desiderata, predictive matching and information consistency. Predictive matching implies that the Bayes factor equals 1 when the data are completely uninformative; information consistency implies that the Bayes factor equals 0 or ∞ when the data are overwhelmingly informative. These desiderata ensure that the correct inference is reached in extreme cases, and in doing so they provide useful restrictions for the specification of the prior distributions.

To achieve the desired result that the Bayes factor equals $\text{BF}_{10}(d) = 1$ for completely uninformative data, π_0, π_1 need to be chosen such that the marginal likelihoods of $\mathcal{M}_0$ and $\mathcal{M}_1$ are predictively matched to each other, that is,

$$\int_{\Theta_0} f(d \mid \theta_0, \mathcal{M}_0)\pi_0(\theta_0)\mathrm{d}\theta_0 = p(d \mid \mathcal{M}_0) = p(d \mid \mathcal{M}_1) = \int_{\Theta_1} f(d \mid \theta_1, \mathcal{M}_1)\pi_1(\theta_1)\mathrm{d}\theta_1 \tag{4}$$

for every completely uninformative data set d.

On the other hand, when data d are overwhelmingly informative in favor of the alternative model the Bayes factor should yield $\text{BF}_{10}(d) = \infty$ or, equivalently, $\text{BF}_{01}(d) = 1/\text{BF}_{10}(d) = 0$, as this then yields $P(\mathcal{M}_1 \mid d) = 1$ for any prior model probability $P(\mathcal{M}_1) > 0$. A Bayes factor with this property is known to be information consistent.

2. Jeffreys's procedure for constructing a default Bayes factor

We now elaborate on Jeffreys's general procedure in constructing default Bayes factors – the specification of the two priors π_0, π_1 – such that the procedure fulfills the desiderata discussed above.

2.1. Step 1. Nest π_0 within π_1

In null hypothesis tests the model $\mathcal{M}_1$ can be considered an extension of $\mathcal{M}_0$ by inclusion of a new parameter, that is, $\theta_1 = (\theta_0, \eta)$ where θ_0 denotes the common parameters and η denotes the test-relevant parameter. Equivalently, $\mathcal{M}_0$ is said to be nested within $\mathcal{M}_1$ due to the connection $f(d \mid \theta_0, \mathcal{M}_0) = f(d \mid \theta_0, \eta = 0, \mathcal{M}_1)$. Jeffreys exploited the connection between these two likelihood functions to induce a relationship between π_1 and π_0. In general one has $\pi_1(\theta_0, \eta) = \pi_1(\eta \mid \theta_0)\pi_1(\theta_0)$, but due to the nesting Jeffreys treats the common parameters within $\mathcal{M}_1$ as in $\mathcal{M}_0$, that is, $\pi_1(\theta_0) = \pi_0(\theta_0)$. Furthermore, when η can be sensibly related to θ_0, Jeffreys redefines the test-relevant parameter as δ, and decomposes the prior as $\pi_1(\eta, \theta_0) = \pi_1(\delta)\pi_0(\theta_0)$. For instance, in the case of the t-test Jeffreys focuses on effect size $\delta = \frac{\mu}{\sigma}$.

This implies that once we have chosen π_0, we have then completely specified the marginal likelihood $p(d \mid \mathcal{M}_0)$ and can, therefore, readily calculate the denominator of the Bayes factor $\text{BF}_{10}(d)$ given data d. Furthermore, due to the nesting of π_0 within π_1 we can also calculate a large part of the marginal likelihood of $\mathcal{M}_1$ as

$$p(d \mid \mathcal{M}_1) = \int_{\Delta} \underbrace{\int_{\Theta} f(d \mid \theta_0, \delta, \mathcal{M}_1)\pi_0(\theta_0)\mathrm{d}\theta_0}_{h(d \mid \delta)} \pi_1(\delta)\mathrm{d}\delta, \tag{5}$$

where $h(d \mid \delta)$ is the test-relevant likelihood, a function that only depends on the data and the test-relevant parameter δ as the common parameters θ_0 are integrated out. The following two steps are concerned with choosing $\pi_1(\delta)$ such that the resulting Bayes factor is well-calibrated to extreme data.

2.2. Step 2. Predictive matching

Typically, a certain minimum number of samples n_{min} is required before model $\mathcal{M}_1$ can be differentiated from $\mathcal{M}_0$. All possible data sets with sample sizes less than n_{min} are considered uninformative. For example, at least $n_{\text{min}} = 2$ observations are required to distinguish $\mathcal{M}_0 : \mu = 0$ from $\mathcal{M}_1$ in a t-test. Specifically, confronted with a single Gaussian observation unequal to zero, for instance, $x_1 = 5$, lack of knowledge about σ within $\mathcal{M}_0$ means that we cannot exclude $\mathcal{M}_0$ as a reasonable explanation for the data.

Indeed, a member of $\mathcal{M}_0$, a zero-mean normal distribution with a standard deviation of seven, produces an observation less than five units away from zero with 53% chance. Similarly, lack of knowledge about σ also means that $\mathcal{M}_1$ cannot be excluded as a reasonable explanation of the data. To convey that – for the purpose of discriminating $\mathcal{M}_0$ from $\mathcal{M}_1$ – nothing is learned from any data set with a sample smaller than n_{min} we choose $\pi_1(\delta)$ such that

$$p(d \mid \mathcal{M}_0) = p(d \mid \mathcal{M}_1) = \int_{\Delta} h(d \mid \delta)\pi_1(\delta)\mathrm{d}\delta \tag{6}$$

for every data set d with a sample size less than n_{min}. In sum, $\pi_1(\delta)$ is chosen such that when the data are completely uninformative, $\text{BF}_{10}(d) = 1$.

2.3. Step 3. Information consistency

Even a limited number of observations may provide overwhelming support for $\mathcal{M}_1$. In the case of the t-test, for instance, the support that an observed non-zero mean provides for $\mathcal{M}_1$ should increase without bound when the observed variance, based on any sample size $n \geq n_{\min}$, goes to zero. Consequently, for data d with a sample size greater or equal to $n_{\min}$ that point undoubtedly to $\mathcal{M}_1$, Jeffreys chose $\pi_1(\delta)$ such that $p(d \mid \mathcal{M}_1)$ diverges to infinity. That is, in order to achieve information consistency $p(d \mid \mathcal{M}_0)$ needs to be bounded and $\pi_1(\delta)$ needs to be chosen such that $p(d \mid \mathcal{M}_1) = \int_\Delta h(d \mid \delta)\pi_1(\delta)\mathrm{d}\delta$ diverges to infinity for overwhelmingly informative data of any size n greater or equal to $n_{\min}$.

2.4. Summary

Jeffreys's procedure to construct a Bayes factor begins with the nesting of π_0 within π_1 and the choice of π_0 is, therefore, the starting point of the method. The specification of π_0 yields $p(d \mid \mathcal{M}_0)$. Next, the test-relevant prior π_1 is chosen such that $p(d \mid \mathcal{M}_1)$ is well-calibrated to extreme data that are either completely uninformative or overwhelmingly informative. Together with π_0, this calibrated test-relevant prior forms the basis for Jeffreys's construction of a Bayes factor.

As a default choice for π_0, Jeffreys used his popular "Jeffreys's prior" on the common parameters θ_0 (Jeffreys, 1946). Derived from the likelihood function $f(d \mid \theta_0, \mathcal{M}_0)$, this default prior is translation invariant, meaning that the same posterior is obtained regardless of how the parameters are represented (e.g., Ly, Marsman, Verhagen, Grasman, & Wagenmakers, submitted for publication). Jeffreys's translation-invariant priors are typically improper, that is, non-normalizable, even though they do lead to proper posteriors for the designs discussed below.

The specification of the test-relevant prior requires special care, as priors that are too wide inevitably reduce the weighted likelihood, resulting in a preference for $\mathcal{H}_0$ regardless of the observed data (Jeffreys–Lindley–Bartlett paradox; Bartlett, 1957; Jeffreys, 1961; Lindley, 1957; Marin & Robert, 2010). Consequently, Jeffreys's translation-invariant prior cannot be used for the test-relevant parameter.

Note that Jeffreys's methodical approach in choosing the two priors π_0, π_1 is fully based on the likelihood functions of the two models that are being compared; the priors do not represent substantive knowledge of the parameters within the model and the resulting procedure can therefore be presented as a reference analysis that may be fine-tuned in the presence of additional information. In the following two sections we illustrate Jeffreys's procedure by discussing the development of the default Bayes factors for two scenarios that are particularly relevant for experimental psychology: testing the nullity of a normal mean and the testing the nullity of a correlation coefficient. Appendix A provides a list of additional Bayes factors that are presented in ToP.

3. Jeffreys's Bayes factor for the test of the nullity of a normal mean: The Bayesian t-test

To develop the Bayesian counterpart of the classical t-test we first characterize the data and discuss how they relate to the unknown parameters within each model in terms of the likelihood functions. By studying the likelihood functions we can justify the nesting of π_0 within π_1 and identify data that are completely uninformative and data that are overwhelmingly informative. The test-relevant prior is then selected based on the desiderata discussed above. We then apply the resulting default Bayes factor to an example data set on cheating and creativity. In addition, we develop the one-sided extension of Jeffreys's t-test, after which we conclude with a short discussion.

3.1. Normal data

For the case at hand, experimental outcomes are assumed to follow a normal distribution characterized by the unknown population mean μ and standard deviation σ. Similarly, the observed data d from a normal distribution can be summarized by two numbers: the observed sample mean $\bar{x}$ and the average sums of squares $s_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$; hence we write $d = (\bar{x}, s_n^2)$. The main difference between the null model $\mathcal{M}_0 : \mu = 0$ and its relaxation $\mathcal{M}_1$ is reflected in the population effect size, which is defined as $\delta = \frac{\mu}{\sigma}$, as σ provides a scale to the problem. This population effect size cannot be observed directly, unlike its sampled scaled version the t-statistic, i.e., $t = \frac{\sqrt{n}\bar{x}}{s_\nu}$, where s_ν refers to the sample standard deviation based on $\nu = n - 1$ degrees of freedom. Extreme data can be characterized by $|t| \to \infty$ or equivalently by $s_n^2 \to 0$ and it is used in the calibration step of the Bayes factor to derive the test-relevant prior. To improve readability we remove the subscript n when we refer to the average sum of squares $s^2 = s_n^2$.

3.2. Step 1. Nesting of π_0 within π_1

3.2.1. Comparing the likelihood functions

A model defines a likelihood that structurally relates how the observed data are linked to the unknown parameters. The point null hypothesis $\mathcal{M}_0$ posits that $\mu = 0$, whereas the alternative hypothesis $\mathcal{M}_1$ relaxes the restriction on μ. Conditioned on the observations $d = (\bar{x}, s^2)$, the likelihood functions of $\mathcal{M}_0$ and $\mathcal{M}_1$ are given by

$$f(d \mid \sigma, \mathcal{M}_0) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\tfrac{n}{2\sigma^2}\left[\bar{x}^2 + s^2\right]\right), \tag{7}$$

$$f(d \mid \mu, \sigma, \mathcal{M}_1) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\tfrac{n}{2\sigma^2}\left[(\bar{x} - \mu)^2 + s^2\right]\right), \tag{8}$$

respectively. Note that $f(d \mid \sigma, \mathcal{M}_0)$ is a function of σ alone, whereas $f(d \mid \mu, \sigma, \mathcal{M}_1)$ depends on two parameters, σ and μ. By the nesting we can set $\pi_1(\mu, \sigma) = \pi_1(\mu \mid \sigma)\pi_0(\sigma)$. Jeffreys removed the scale from the problem by considering $\delta = \frac{\mu}{\sigma}$ as the test-relevant parameter which leads to $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$ with a likelihood expressed as

$$f(d \mid \delta, \sigma, \mathcal{M}_1) = (2\pi)^{-\frac{n}{2}} \int_0^\infty \sigma^{-n-1} \times \int_{-\infty}^\infty \exp\left(-\tfrac{n}{2}\left[\left(\tfrac{\bar{x}}{\sigma} - \delta\right)^2 + \left(\tfrac{s}{\sigma}\right)^2\right]\right) \pi_1(\delta)\,\mathrm{d}\delta\,\mathrm{d}\sigma. \tag{9}$$

3.2.2. The denominator of $BF_{10}(d)$

Jeffreys's default choice leads to $\pi_0(\sigma) \propto 1/\sigma$, the translation-invariant prior that Jeffreys's would use to arrive at a posterior for σ within either model. This prior specification leads to the following marginal likelihood of $\mathcal{M}_0$:

$$p(d \mid \mathcal{M}_0) = \begin{cases} \dfrac{1}{2|\bar{x}|} & n = 1, \quad \text{(a)} \\ \dfrac{\Gamma(\frac{n}{2})}{2(\pi n \bar{x}^2)^{\frac{n}{2}}} & n > 1 \text{ and } s^2 = 0, \quad \text{(b)} \\ \dfrac{\Gamma\left(\frac{n}{2}\right)}{2\left(\pi n s^2\right)^{\frac{n}{2}}}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} & n > 1 \text{ and } s^2 > 0, \quad \text{(c)} \end{cases} \tag{10}$$

where t is the observed t-value and ν the degrees of freedom defined as before. Hence, Eq. (10)(a)–(c) define the denominator of the Bayes factor $\mathrm{BF}_{10}(d)$; Eq. (10)(a) will be used to calibrate the Bayes factor $\mathrm{BF}_{10}(d)$ to completely uninformative data, whereas Eq. (10)(b) will be used for the calibration to overwhelmingly informative data. Some statisticians only report the right term

$\left(1+\frac{t^2}{\nu}\right)^{-\frac{n}{2}}$ of Eq. (10)(c), as the first term also appears in the marginal likelihood of $\mathcal{M}_1$ and, thus, cancels out in the Bayes factor.

3.3. Step 2. Predictive matching: Symmetric $\pi_1(\delta)$

We now discuss how the test-relevant prior $\pi_1(\delta)$ can be chosen such that the resulting Bayes factor is well-calibrated. As elaborated above, we consider data sets with only one sample as completely uninformative in discriminating $\mathcal{M}_0$ from $\mathcal{M}_1$. Jeffreys (1961, p. 269) studied Eq. (9) with $n = 1, \bar{x} > 0$, and, consequently, $s^2 = 0$, and concluded that $p(d \mid \mathcal{M}_1)$ is matched to Eq. (10)(a) whenever $\pi_1(\delta)$ is symmetric around zero.

3.4. Step 3. Information consistency: Heavy-tailed $\pi_1(\delta)$

On the other hand, observed data $\bar{x} > 0, s^2 = 0$ with $n > 1$ can be considered overwhelmingly informative as the t-value is then infinite. To obtain maximum evidence in favor of the alternative we require that $\text{BF}_{10}(d) = \infty$. This occurs whenever the marginal likelihood of $\mathcal{M}_1$ is infinite and $p(d \mid \mathcal{M}_0)$ finite, see Eq. (10)(b). Jeffreys (1961, p. 269–270) showed that this is the case whenever the test-relevant prior $\pi_1(\delta)$ is heavy-tailed.

3.5. The resulting Bayes factor

Hence, a Bayes factor that meets Jeffreys's desiderata can be obtained by assigning $\pi_0(\sigma) \propto 1/\sigma$ and $\pi_1(\delta, \sigma) = \pi_1(\delta)\pi_0(\sigma)$, where $\pi_1(\delta)$ is symmetric around zero and heavy-tailed.

3.5.1. Jeffreys's choice: The standard Cauchy distribution

The Cauchy distribution with scale γ is the most well-known distribution which is both symmetric around zero and heavy-tailed:

$$\pi_1(\delta\,;\,\gamma) = \frac{1}{\pi\gamma\left(1+\left(\frac{\delta}{\gamma}\right)^2\right)}. \tag{11}$$

As a default choice for $\pi_1(\delta)$, Jeffreys suggested to use the simplest version, the standard Cauchy distribution with $\gamma = 1$.

3.5.2. Jeffreys's Bayesian t-test

Jeffreys's Bayes factor now follows from the integral in Eq. (9) with respect to Cauchy distributions $\pi_1(\delta)$ divided by Eq. (10)(c), whenever $n > 1$ and $s^2 > 0$. Jeffreys knew that this integral is hard to compute and went to great lengths to compute an approximation that makes his Bayesian t-test usable in practice. Fortunately, we can now take advantage of computer software that can numerically solve the aforementioned integral and we therefore omit Jeffreys's approximation from further discussion. By a decomposition of a Cauchy distribution we obtain a Bayes factor of the following form:

$$\text{BF}_{10\,;\,\gamma}(n,t) = \frac{\gamma\int_0^\infty (1+ng)^{-\frac{1}{2}}\left(1+\frac{t^2}{\nu(1+ng)}\right)^{-\frac{\nu+1}{2}}(2\pi)^{-\frac{1}{2}}g^{-\frac{3}{2}}e^{-\frac{\gamma^2}{2g}}\,\mathrm{d}g}{\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}}, \tag{12}$$

where g is an auxiliary variable that is integrated out numerically. Jeffreys's choice is obtained when $\gamma = 1$. The Bayes factor $\text{BF}_{10\,;\,\gamma=1}(n,t)$ now awaits a user's observed t-value and the associated n number of observations.

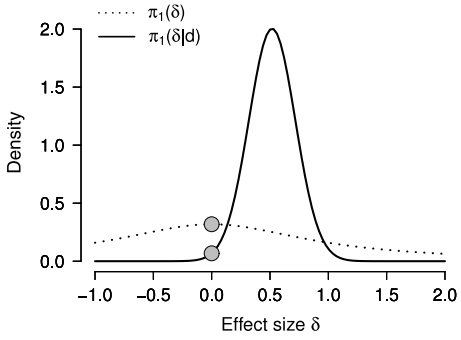

Fig. 1. Posterior and prior distributions of the effect size for a two-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $\text{BF}_{10\,;\,\gamma=1} \approx 4.60$ equals the height ratio of the prior distribution $\pi_1(\delta)$ over the posterior distribution $\pi_1(\delta \mid d)$ at $\delta = 0$.

3.6. Example: The Bayesian between-subject t-test

To illustrate the default Bayesian t-test we extend Eq. (12) to a between-subjects design and apply the test to a psychological data set. The development above is easily generalized to a between-subject design in which observations are assumed to be drawn from two separate normal populations. To do so, we replace: (i) the value of t by the observed two-sample (grouped) t value, (ii) the effective sample size by $n = \frac{n_1 n_x}{n_1+n_2}$, and (iii) the degrees of freedom with $\nu = n_1 + n_2 - 2$, see Rouder et al. (2009).

Example 1 (*Does Cheating Enhance Creativity?*)**.** Gino and Wiltermuth (2014, Experiment 2) reported that the act of cheating enhances creativity. This conclusion was based on five experiments. Here we analyze the results from Experiment 2 in which, having been assigned either to a control condition or to a condition in which they were likely to cheat, participants were rewarded for correctly solving each of 20 math and logic multiple-choice problems. Next, participants' creativity levels were measured by having them complete 12 problems from the Remote Association Task (RAT; Mednick, 1962).

The control group featured $n_1 = 48$ participants who scored an average of $\bar{x}_1 = 4.65$ RAT items correctly with a sample standard deviation of $s_{n_1-1} = 2.72$. The cheating group featured $n_2 = 51$ participants who scored $\bar{x}_2 = 6.20$ RAT items correctly with $s_{n_2-1} = 2.98$. These findings yield $t(97) = 2.73$ with $p = .008$. Jeffreys's default Bayes factor yields $\text{BF}_{10}(d) \approx 4.6$, indicating that the data are 4.6 times more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$. With equal prior odds, the posterior probability for $\mathcal{M}_0$ remains an arguably non-negligible 17%.

For nested models, the Bayes factor can be obtained without explicit integration, using the Savage–Dickey density ratio test (e.g., Dickey & Lientz, 1970; Marin & Robert, 2010; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). The Savage–Dickey test is based on the following identify:

$$\text{BF}_{10}(d) = \frac{\pi_1(\delta = 0)}{\pi_1(\delta = 0 \mid d)}. \tag{13}$$

One of the additional advantages of the Savage–Dickey test is that it allows the result of the test to be displayed visually, as the height of the prior versus the posterior at the point of test (i.e., $\delta = 0$). Fig. 1 presents the results from Experiment 2 of Gino and Wiltermuth (2014). ◇

In this example, both the Bayesian and Fisherian analysis gave the same qualitative result. Nevertheless, the Bayes factor is more conservative, and some researchers may be surprised that, for the same data, $p = .008$ and posterior model probability $P(\mathcal{M}_0 \mid d) = .17$. Indeed, for many cases the Bayesian and Fisherian analyses disagree qualitatively as well as quantitatively (e.g., Wetzels et al., 2011).

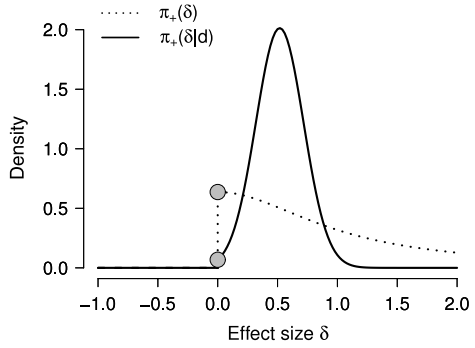

Fig. 2. Posterior and prior distributions of effect size for a one-sided default Bayes factor analysis of Experiment 2 of Gino and Wiltermuth (2014). The Jeffreys default Bayes factor of $\text{BF}_{+0} = 9.18$ equals the height ratio of the prior distribution $\pi_1(\delta)$ over the posterior distribution $\pi_1(\delta \mid d)$ at $\delta = 0$. The prior distribution $\pi_+(\delta)$ is zero for negative values of δ. Furthermore, note that the prior distribution for $\delta \geq 0$ is twice as high compared to $\pi_1(\delta)$ in Fig. 1.

3.7. The one-sided extension of Jeffreys's Bayes factor

Some reflection suggests that the authors' hypothesis from Example 1 is more specific—the authors argued that cheating leads to more creativity, not less. To take into account the directionality of the hypothesis we need a one-sided adaptation of Jeffreys's Bayes factor $\text{BF}_{10\,;\,\gamma=1}(n, t)$. The comparison that is made is then between the model of no effect $\mathcal{M}_0$ and one denoted by $\mathcal{M}_+$ in which the effect size δ is assumed to be positive. We decompose $\text{BF}_{+0}(d)$ as follows:

$$\text{BF}_{+0}(d) = \underbrace{\frac{p(d \mid \mathcal{M}_+)}{p(d \mid \mathcal{M}_1)}}_{\text{BF}_{+1}(d)} \underbrace{\frac{p(d \mid \mathcal{M}_1)}{p(d \mid \mathcal{M}_0)}}_{\text{BF}_{10}(d)}, \tag{14}$$

where $\text{BF}_{+1}(d)$ is the Bayes factor that compares the unconstrained model $\mathcal{M}_1$ to the positively restricted model $\mathcal{M}_+$ (Morey & Wagenmakers, 2014; Mulder, Hoijtink, & Klugkist, 2010; Pericchi, Liu, & Torres, 2008). The objective comparison between $\mathcal{M}_+$ and $\mathcal{M}_1$ is then to keep all aspects the same: $\pi_+(\sigma) = \pi_1(\sigma) = \pi_0(\sigma)$ except for the distinguishing factor of δ being restricted to positive values within $\mathcal{M}_+$. For the test-relevant prior distribution we restrict $\pi_1(\delta)$ to positive values of δ, which by symmetry of the Cauchy distribution means that $\pi_+(\delta)$ accounts doubly for the likelihood when δ is positive and nullifies it when δ is negative (Klugkist, Laudy, & Hoijtink, 2005).

Example 1 (*One-Sided Test for the Gino and Wiltermuth Data, continues*)**.** For the data from Gino and Wiltermuth (2014, Experiment 2) the one-sided adaptation of Jeffreys's Bayes factor equation (12) yields $\text{BF}_{+0}(d) = 9.18$. Because almost all of the posterior mass is consistent with the authors' hypothesis, the one-sided Bayes factor is almost twice the two-sided Bayes factor. The result is visualized through the Savage–Dickey ratio in Fig. 2. ◇

3.8. Discussion on the t-test

In this section we showcased Jeffreys's procedure in selecting the instrumental priors π_0, π_1 that yield a Bayes factor for grading the support that the data provide for $\mathcal{M}_0$ versus $\mathcal{M}_1$. The construction of this Bayes factor began by assigning Jeffreys's translation-invariant prior to the common parameters, that is, $\pi_0(\sigma) \propto 1/\sigma$. This is the same prior Jeffreys would use for estimating σ in either of the two models, when one of these two models is assumed to hold true. This prior on the common parameters then yields the denominator of the Bayes factor Eq. (10)(a)–(c). Jeffreys noted that when the test-relevant prior $\pi_1(\delta)$ is symmetric and heavy tailed, the resulting Bayes factor is guaranteed to yield the correct conclusion for completely uninformative data and for overwhelmingly informative data. Jeffreys (1961, p. 272–273) noted that the standard Cauchy prior for δ yields a Bayes factor equation (12) (with $\gamma = 1$) that aligns with this calibration.

It took several decades before Jeffreys's Bayes factor for the t-test was adopted by Zellner and Siow (1980) who generalized it to the linear regression framework based on a multivariate Cauchy distribution. One practical drawback of their proposal was the fact that the numerical integration required to calculate the Bayes factor becomes computationally demanding as the number of covariates grows.

Liang, Paulo, Molina, Clyde, and Berger (2008) proposed a computationally efficient alternative to the Zellner and Siow (1980) setup by first decomposing the multivariate Cauchy distribution into a mixture of gamma and normal distributions followed by computational simplifications introduced by Zellner (1986). As a result, the Bayes factor can be obtained from only a single numerical integral, regardless of the number of covariates. The form of the numerator in Eq. (12) is in fact inspired by Liang et al. (2008) and introduced to psychology by Rouder et al. (2009) and Wetzels, Raaijmakers, Jakab, and Wagenmakers (2009). The combination $\pi_0(\sigma) \propto \sigma^{-1}$ and $\delta \sim \mathcal{C}(0, 1)$ was dubbed the JZS-prior in honor of Jeffreys, Zellner, and Siow; this is understandable in the framework of linear regression, although it should be noted that all ideas for the t-test were already present in the second edition of ToP (Jeffreys, 1948, p. 242–248).

3.8.1. Model selection consistency

In addition to predictive matching and information consistency, Liang et al. (2008) showed that Zellner and Siow's 1980 generalization of Jeffreys's work is also model selection consistent, which implies that as the sample size n increases indefinitely, the support that the data d provide for the correct data-generating model (i.e., $\mathcal{M}_0$ or $\mathcal{M}_1$) grows without bound. Hence, Jeffreys's default Bayes factor equation (12) leads to the correct decision whenever the sample size is sufficiently large. Jeffreys's procedure of assigning default priors for Bayesian hypothesis testing was recently generalized by Bayarri, Berger, Forte, and García-Donato (2012). We now turn to Jeffreys's development of another default Bayes factor: the test for the presence of a correlation.

4. Jeffreys's Bayes factor for the test of the nullity of a correlation

To develop the Bayesian correlation test we first characterize the data and discuss how they relate to the unknown parameters within each model in terms of the likelihood functions. By studying the likelihood functions we can justify the nesting of π_0 within π_1 and identify data that are completely uninformative and data that are overwhelmingly informative. As was done for the Bayesian t-test, the test-relevant prior is selected based on a calibration argument. The derivations and calibrations given here cannot be found in Jeffreys (1961), as Jeffreys appears to have derived the priors intuitively. Hence, we divert from the narrative of Jeffreys (1961, Paragraph 5.5) and instead: (a) explain Jeffreys's reasoning with a structure analogous to that of the previous section; and (b) give the exact results instead, as Jeffreys used an approximation to simplify the calculations. In effect, we show that Jeffreys's intuitive choice is very close to our exact result. After presenting the correlation Bayes factor we relate it to Jeffreys's choice and apply the resulting default Bayes factor to an example data set that is concerned with presidential height and the popular vote. In addition, we develop the one-sided extension of Jeffreys's correlation test, after which we conclude with a short discussion.

4.1. Bivariate normal data

For the case at hand, experimental outcome pairs (X, Y) are assumed to follow a bivariate normal distribution characterized by the unknown population means μ_x, μ_y, standard deviations σ, υ of X and Y respectively. Within $\mathcal{M}_1$ the parameter ρ characterizes the linear association between X and Y. To test the nullity of the population correlation it is helpful to summarize the data for X and Y separately in terms of their respective sample means and average sums of squares: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $u^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, respectively. The sample correlation coefficient r then defines an observable measure of the linear relationship between X and Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{-n\bar{x}\bar{y} + \sum_{i=1}^{n} x_i y_i}{nsu}. \tag{15}$$

This sample correlation coefficient r is an imperfect reflection of the unobservable population correlation coefficient ρ. The data can be summarized by the five quantities $d = (\bar{x}, s^2, \bar{y}, u^2, r)$.

The main difference between the null model $\mathcal{M}_0$ and $\mathcal{M}_1$ is reflected in the population correlation coefficient ρ, which cannot be observed directly, unlike its sampled version known as Pearson's r, Eq. (15). Extreme data can be characterized by $|r| = 1$ and this is used in the calibration step of the Bayes factor to derive the form of the test-relevant prior.

4.2. Step 1. Nesting of π_0 within π_1

4.2.1. Comparing the likelihood functions

The point null hypothesis $\mathcal{M}_0$ assumes that the data follow a bivariate normal distribution with ρ known and fixed at zero. Hence, $\mathcal{M}_0$ depends on four parameters which we abbreviate as $\theta_0 = (\mu_x, \mu_y, \sigma, \upsilon)$, while the alternative model $\mathcal{M}_1$ can be considered an extension of $\mathcal{M}_0$ with an additional parameter ρ, i.e., $\theta_1 = (\theta_0, \rho)$. These two bivariate normal models relate the observed data to the parameters using the following two likelihood functions:

$$\begin{aligned} f(d \mid \theta_0, \mathcal{M}_0) &= (2\pi\sigma\upsilon)^{-n} \exp\left(-\tfrac{n}{2}\left[\left(\tfrac{\bar{x}-\mu_x}{\sigma}\right)^2 + \left(\tfrac{\bar{y}-\mu_y}{\upsilon}\right)^2\right]\right) \\ &\quad \times \exp\left(-\tfrac{n}{2}\left[\left(\tfrac{s}{\sigma}\right)^2 + \left(\tfrac{u}{\upsilon}\right)^2\right]\right). \end{aligned} \tag{16}$$

$$\begin{aligned} f(d \mid \theta_1, \mathcal{M}_1) &= \left(2\pi\sigma\upsilon\sqrt{1-\rho^2}\right)^{-n} \exp\left(-\tfrac{n}{2(1-\rho^2)}\left[\frac{(\bar{x}-\mu_x)^2}{\sigma^2}\right.\right. \\ &\quad \left.\left. - 2\rho\frac{(\bar{x}-\mu_x)(\bar{y}-\mu_y)}{\sigma\upsilon} + \frac{(\bar{y}-\mu_y)^2}{\upsilon^2}\right]\right) \\ &\quad \times \exp\left(-\frac{n}{2(1-\rho^2)}\left[\left(\tfrac{s}{\sigma}\right)^2 - 2\rho\left(\tfrac{rsu}{\sigma\upsilon}\right) + \left(\tfrac{u}{\upsilon}\right)^2\right]\right). \end{aligned} \tag{17}$$

Note that $f(d \mid \theta_0, \mathcal{M}_0) = f(d \mid \theta_0, \rho = 0, \mathcal{M}_1)$ and because the population correlation ρ is defined as

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{E(XY) - \mu_x\mu_y}{\sigma\upsilon}, \tag{18}$$

we know that ρ remains the same under data transformations of the form $\tilde{X} = aX - b$, $\tilde{Y} = cY - d$. In particular, we can take $b = \mu_x$, $d = \mu_y$, $a = 1/\sigma$, $c = 1/\upsilon$ and conclude that ρ does not depend on these common parameters θ_0. Hence, we nest π_0 within π_1 orthogonally, that is, $\pi_1(\theta_0, \rho) = \pi_1(\rho)\pi_0(\theta_0)$.

4.2.2. The denominator of $BF_{10}(d)$

Jeffreys's default choice leads to assigning $\pi_0(\theta_0)$ the joint prior $\pi_0(\mu_x, \mu_y, \sigma, \upsilon) = 1 \cdot 1 \cdot \frac{1}{\sigma}\frac{1}{\upsilon}$; this is the translation-invariant prior that Jeffreys would use to update to the posterior for θ_0 within either model. When the averaged sum of squares are both non-zero, this yields the following marginal likelihood of $\mathcal{M}_0$:

$$p(d \mid \mathcal{M}_0) = 2^{-2} n^{-n} \pi^{1-n} (su)^{1-n} \left[\Gamma\left(\tfrac{n-1}{2}\right)\right]^2. \tag{19}$$

Eq. (19) defines the denominator of the correlation Bayes factor $\mathrm{BF}_{10}(d)$. Observe that this marginal likelihood does not depend on the sample correlation coefficient r.

4.3. Step 2. Predictive matching: Symmetric $\pi_1(\rho)$

4.3.1. Deriving the test-relevant likelihood function

We now discuss how the test-relevant prior $\pi_1(\rho)$ can be defined such that the resulting Bayes factor is well-calibrated. The conclusion is as before: we require $\pi_1(\rho)$ to be symmetric around zero. We discuss the result more extensively as it cannot be found in Jeffreys (1961). Furthermore, the test-relevant likelihood function reported by Jeffreys (1961, p. 291, Eq. 8) is in fact an approximation of the result given below.

Before we can discuss the calibration we first derive the test-relevant likelihood function by integrating out the common parameters θ_0 from Eq. (17) with respect to the translation-invariant priors $\pi_0(\theta_0)$ as outlined by Eq. (5). This leads to the following simplification:

$$p(d \mid \mathcal{M}_1) = p(d \mid \mathcal{M}_0) \int_{-1}^{1} h(n, r \mid \rho)\pi_1(\rho)\mathrm{d}\rho, \tag{20}$$

where h is the test-relevant likelihood function that depends on n, r, ρ alone and is given by Eqs. (22) and (23). The Bayes factor, therefore, reduces to

$$\begin{aligned} \mathrm{BF}_{10}(d) &= \frac{p(d \mid \mathcal{M}_1)}{p(d \mid \mathcal{M}_0)} = \frac{p(d \mid \mathcal{M}_0)\int_{-1}^{1} h(n, r \mid \rho)\pi_1(\rho)\mathrm{d}\rho}{p(d \mid \mathcal{M}_0)} \\ &= \int_{-1}^{1} h(n, r \mid \rho)\pi_1(\rho)\mathrm{d}\rho. \end{aligned} \tag{21}$$

Note that whereas $p(d \mid \mathcal{M}_0)$ does not depend on ρ or the statistic r (see Eq. (19)), the function h does not depend on the statistics $\bar{x}, s^2, \bar{y}, u^2$ that are associated with the common parameters. Thus, the evidence for $\mathcal{M}_1$ over $\mathcal{M}_0$ resides within n and r alone.

The test-relevant likelihood function $h(n, r \mid \rho)$ possess more regularities. In particular, it can be decomposed into an even and an odd function, that is, $h = A + B$, with A defined as

$$A(n, r \mid \rho) = (1-\rho^2)^{\frac{n-1}{2}}\, {}_2F_1\left(\tfrac{n-1}{2}, \tfrac{n-1}{2}; \tfrac{1}{2}; (r\rho)^2\right), \tag{22}$$

where ${}_2F_1$ is Gauss' hypergeometric function (see Appendix B for details). Observe that A is a symmetric function of ρ when n and r are given. The second function B is relevant for the one-sided test and is given by

$$B(n, r \mid \rho) = 2r\rho(1-\rho^2)^{\frac{n-1}{2}} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right]^2 {}_2F_1\left(\tfrac{n}{2}, \tfrac{n}{2}; \tfrac{3}{2}; (r\rho)^2\right), \tag{23}$$

which is an odd function of ρ when n and r are given. Thus, the test-relevant likelihood function h that mediates inference about the presence of ρ from n and r is given by $h(n, r \mid \rho) = A(n, r \mid \rho) + B(n, r \mid \rho)$. Examples of the functions A and B are shown in Fig. 3.

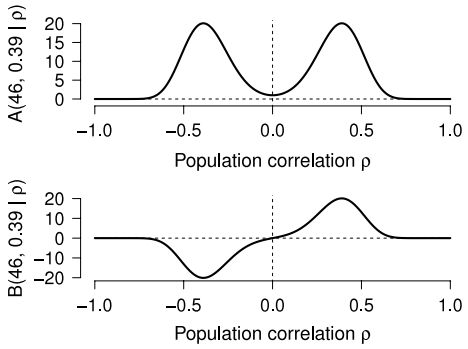

Fig. 3. $A(n, r \mid \rho)$ is an even function of ρ, and $B(n, r \mid \rho)$ is an odd function of ρ. Together, A and B determine the function h from Eq. (21): $h(n, r \mid \rho) = A(n, r \mid \rho) + B(n, r \mid \rho)$. For this illustration, we used $n = 46$ and $r = 0.39$ based on the example data discussed below.

4.3.2. Predictive matching and the minimal sample size of $n_{\min} = 3$

Interestingly, the predictive matching principle implies the use of a symmetric test-relevant prior as in the previous case. Note that we cannot infer the correlation of a bivariate normal distribution whenever we have only a single data pair (x, y); r is undefined when $n = 1$. Furthermore, when $n = 2$ we automatically get $r = 1$ or $r = -1$ regardless of whether or not $\rho = 0$ is true. As such, nothing is learned up to $n_{\min} = 3$ when testing the nullity of ρ. Hence, we have to choose $\pi_1(\rho)$ such that the resulting Bayes factor equation (21) equals one for $n = 1$ and $n = 2$ regardless of the actually observed r.

Using $n = 1$ in Eqs. (22) and (23) we see that $h(1, r \mid \rho) = A(1, r \mid \rho) + B(1, r \mid \rho) = 1$ for every ρ and r. From a consideration of Eq. (21) it follows that for a Bayes factor of one with $n = 1$, we require $\pi_1(\rho)$ to integrate to one (i.e., $\text{BF}_{10}(d) = \int_{-1}^{1} \pi_1(\rho)\text{d}\rho = 1$), underscoring Jeffreys's claim that test-relevant priors should be proper.[1] Similarly, for $n = 2$ we automatically obtain $|r| = 1$ and plugging this into Eq. (22) yields $A(2, |r| = 1 \mid \rho) = 1$. Thus, with $\pi_1(\rho)$ a proper prior this yields a Bayes factor of $\text{BF}_{10}(d) = 1 + \int_{-1}^{1} B(2, |r| = 1 \mid \rho)\pi_1(\rho)\text{d}\rho$. To ensure that the Bayes factor equals one for data with a sample size of $n = 2$ we have to nullify the contribution of the function $B(2, |r| = 1 \mid \rho)$. This occurs when $\pi_1(\rho)$ is symmetric around zero, since $B(2, r \mid \rho)$ is an odd function of ρ, see Fig. 3.

4.4. Step 3. Information consistency

On the other hand, a sample correlation $r = 1$ or $r = -1$ with $n \geq n_{\min} = 3$ can be considered overwhelmingly informative data in favor of the alternative model $\mathcal{M}_1$. In our quest to find the right test-relevant prior that yields a Bayes factor that is information consistent, we consider the so-called stretched symmetric beta distributions given by

$$\pi_1(\rho\,;\,\kappa) = \frac{2^{\frac{\kappa-2}{\kappa}}}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa})}(1-\rho^2)^{\frac{1-\kappa}{\kappa}}, \tag{24}$$

where $\mathcal{B}(1/\kappa, 1/\kappa)$ is a beta function, see Appendix C for details. Each $\kappa > 0$ yields a candidate test-relevant prior. Jeffreys's intuitive choice is represented by Eq. (24) with $\kappa = 1$, as this choice corresponds to the uniform distribution of ρ on $(-1, 1)$. Furthermore, κ can be thought of as a scale parameter of the prior as in Eq. (11). We claim that a Bayes factor based on a test-relevant prior Eq. (24) with $\kappa \geq 2$ is information consistent.

[1] Jeffreys rejected the translation-invariant prior $\rho \propto (1-\rho^2)^{-1}$ because it leads to unwelcome results when testing the null hypothesis $\rho = 1$. However, Robert et al. (2009) noted that such a test is rather uncommon as interest typically centers on the point null hypothesis $\mathcal{M}_0 : \rho = 0$.

Table 1
The Bayes factor $\text{BF}_{10\,;\,\kappa=2}$ is information consistent as it diverts to infinity when $r = 1$ and $n \geq 3$, while Jeffreys's intuitive choice $\text{BF}_{10\,;\,\kappa=1}$ does not do so until $n \geq 4$. Hence, Jeffreys intuitive choice $\kappa = 1$ misses the information consistency criterion by one observation. Furthermore, note the role of κ; the smaller it is, the stronger the associated Bayes factors violate the criterion of information consistency.

n	$\text{BF}_{10\,;\,\kappa=5}$	$\text{BF}_{10\,;\,\kappa=2}$	$\text{BF}_{10\,;\,\kappa=1}$	$\text{BF}_{10\,;\,\kappa=1/3}$	$\text{BF}_{10\,;\,\kappa=1/10}$
1	1	1	1	1	1
2	1	1	1	1	1
3	∞	∞	2	1.2	1.05
4	∞	∞	∞	1.75	1.17
5	∞	∞	∞	3.20	1.36

4.5. The resulting Bayes factor

To prove the information consistency claim, ρ is integrated out of the test-relevant likelihood with $h = A + B$ as discussed above (Eq. (21)). This results in the following analytical Bayes factor:

$$\begin{aligned}\text{BF}_{10\,;\,\kappa}(n, r) &= \int_{-1}^{1} h(n, r \mid \rho)\pi_1(\rho\,;\,\kappa)\text{d}\rho \\ &= \int_{-1}^{1} A(n, r \mid \rho)\pi(\rho\,;\,\kappa)\text{d}\rho + \underbrace{\int_{-1}^{1} B(n, r \mid \rho)\pi(\rho\,;\,\kappa)\text{d}\rho}_{0} \\ &= \frac{2^{\frac{\kappa-2}{\kappa}}\sqrt{\pi}}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa})}\frac{\Gamma\left(\frac{2+(n-1)\kappa}{2\kappa}\right)}{\Gamma\left(\frac{2+n\kappa}{2\kappa}\right)}\,{}_2F_1\left(\tfrac{n-1}{2}, \tfrac{n-1}{2}\,;\,\tfrac{2+n\kappa}{2\kappa}\,;\,r^2\right),\end{aligned} \tag{25}$$

where the contribution of the B-function is nullified due to symmetry of the prior. We call Eq. (25) Jeffreys's exact correlation test, as we believe that Jeffreys would have derived this Bayes factor $\text{BF}_{10\,;\,\kappa}(n, r)$, if he had deemed it necessary to calculate it exactly.

Table 1 lists the Bayes factors for a selection of values for κ and n with $r = 1$ fixed; the results confirm that the Bayes factor is indeed information consistent when $\kappa \geq 2$. Note that Jeffreys's choice of $\kappa = 1$ does not lead to a Bayes factor which provides extreme support for $\mathcal{M}_1$ when confronted with data that are overwhelmingly informative (i.e., $r = 1$ and $n_{\min} = 3$). However, this Bayes factor does diverge when $n \geq 4$. Thus, Jeffreys's intuitive choice for κ misses the information consistency criterion by one data pair. The resulting Bayes factor $\text{BF}_{10\,;\,\kappa}(n, r)$ now awaits a user's observed r-value and the associated n number of observations. In what follows, we honor Jeffreys's intuition and showcase the correlation Bayes factor using Jeffreys's choice $\kappa = 1$.

4.6. Example: The Bayesian correlation test

We now apply Jeffreys's default Bayesian correlation test to a data set analyzed earlier by Stulp, Buunk, Verhulst, and Pollet (2013).

Example 2 (*Do Taller Electoral Candidates Attract More Votes?*)**.** Stulp et al. (2013) studied whether there exists a relation between the height of electoral candidates and their popularity among voters. Based on the data from $n = 46$ US presidential elections, Stulp et al. (2013) reported a positive linear correlation of $r = .39$ between X, the relative height of US presidents compared to their opponents, and Y, the proportion of the popular vote. A frequentist analysis yielded $p = .007$. Fig. 4 displays the data. Based in part on these results, Stulp et al. (2013, p. 159) concluded that "height is indeed an important factor in the US presidential elections", and "The advantage of taller candidates is potentially explained by perceptions associated with height: taller presidents are rated by experts as 'greater', and having more leadership and

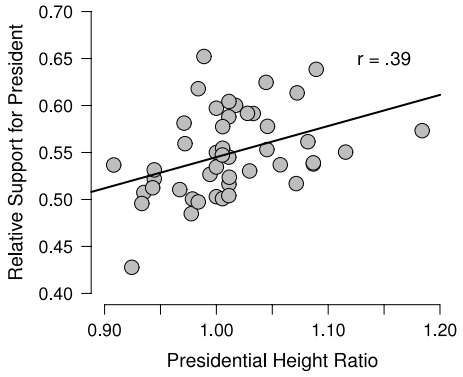

Fig. 4. The data from $n = 46$ US presidential elections, showing the proportion of the popular vote for the president versus his relative height advantage against the closest competitor. The sample correlation equals $r = .39$, and, assuming an unrealistic sampling plan, the p-value equals .007. Jeffreys's default two-sided Bayes factor equals $\text{BF}_{10;\kappa=1}(n = 46, r = .39) = 6.33$, and the corresponding one-sided Bayes factor equals $\text{BF}_{+0;\kappa=1}(n = 46, r = .39) = 11.87$. See text for details.

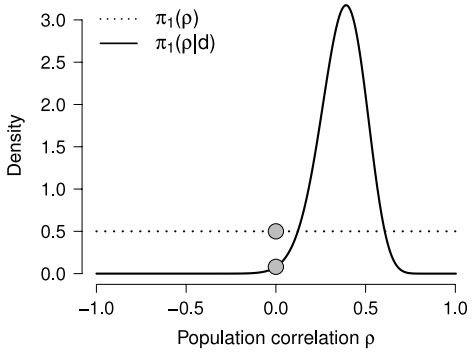

Fig. 5. Posterior and prior distributions of the population correlation coefficient ρ for a two-sided default Bayes factor analysis of the height-popularity relation in US presidents (Stulp et al., 2013). The Jeffreys default Bayes factor of $\text{BF}_{10;\kappa=1} = 6.33$ equals the height ratio of the prior distribution $\pi_1(\rho)$ over the posterior distribution $\pi_1(\rho \mid d)$ at $\rho = 0$.

communication skills. We conclude that height is an important characteristic in choosing and evaluating political leaders".

For the Stulp et al. (2013) election data Jeffreys's exact correlation Bayes factor equation (25) yields $\text{BF}_{10;\kappa=1} = 6.33$, indicating that the observed data are 6.33 times more likely under $\mathcal{M}_1$ than under $\mathcal{M}_0$. This result is visualized in Fig. 5 using the Savage–Dickey density ratio test. With equal prior odds, the posterior probability for $\mathcal{M}_0$ remains an arguably non-negligible 14%. ◇

4.7. The One-sided extension of Jeffreys's exact correlation Bayes factor

Whereas the function A fully determines the two-sided Bayes factor $\text{BF}_{10;\kappa}(n, r)$, the function B takes on a prominent role when we compare the null hypothesis $\mathcal{M}_0$ against the one-sided alternative $\mathcal{M}_+$ with $\rho > 0$.

To extend Jeffreys's exact correlation Bayes factor to a one-sided version, we retain the prior on the common parameters θ_0. For the test-relevant prior $\pi_+(\rho \mid \kappa)$ we restrict ρ to non-negative values, which due to symmetry of $\pi_1(\rho \mid \kappa)$ is specified as

$$\pi_+(\rho\,;\,\kappa) = \begin{cases} 2\pi_1(\rho\,;\,\kappa) & \text{for } 0 \leq \rho \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Recall that A is an even function of ρ; combined with the doubling of the prior for ρ this leads to a one-sided Bayes factor that can be

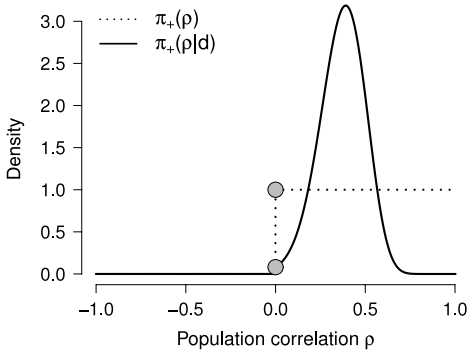

Fig. 6. Posterior and prior distributions of the population correlation coefficient ρ for a one-sided default Bayes factor analysis of the height-popularity relation in US presidents (Stulp et al., 2013). The Jeffreys default Bayes factor of $\text{BF}_{+0;\kappa=1} = 11.87$ equals the height ratio of the prior $\pi_+(\rho)$ over the posterior $\pi_+(\rho \mid d)$ at $\rho = 0$. The prior $\pi_+(\rho)$ is zero for negative values of ρ. Furthermore, note that the prior distribution $\pi_+(\rho)$ is twice as high for $\rho \geq 0$ compared to $\pi_1(\rho)$ in Fig. 5.

decomposed as

$$\text{BF}_{+0;\kappa}(n, r) = \underbrace{\text{BF}_{10;\kappa}(n, r)}_{\int_0^1 A(n,r \mid \rho)\pi_+(\rho\,;\,\kappa)\mathrm{d}\rho} + \underbrace{C_{+0;\kappa}(n, r)}_{\int_0^1 B(n,r \mid \rho)\pi_+(\rho\,;\,\kappa)\mathrm{d}\rho}. \tag{27}$$

The function $C_{+0;\kappa}(n, r)$ can be written as

$$C_{+0;\kappa}(n, r) = \frac{2^{\frac{3\kappa-2}{\kappa}} r\kappa}{\mathcal{B}(\frac{1}{\kappa}, \frac{1}{\kappa})\Big((n-1)\kappa + 2\Big)} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right]^2 \times {}_3F_2\left(1, \tfrac{n}{2}, \tfrac{n}{2}\,;\, \tfrac{3}{2}, \tfrac{2+\kappa(n+1)}{\kappa}\,;\, r^2\right), \tag{28}$$

where ${}_3F_2$ is a generalized hypergeometric function (Gradshteyn & Ryzhik, 2007, p. 1010) with three upper and two lower parameters.

The function $C_{+0;\kappa}(n, r)$ is positive whenever r is positive, since B as a function of ρ is then positive on the interval $(0, 1)$; consequently, for positive values of r the restricted, one-sided alternative hypothesis M_+ is supported more than the unrestricted, two-sided hypothesis $\mathcal{M}_1$, that is, $\text{BF}_{+0;\kappa}(n, r) > \text{BF}_{10;\kappa}(n, r)$. On the other hand, $C_{+0;\kappa}(n, r)$ is negative whenever r is negative; for such cases, $\text{BF}_{+0;\kappa}(n, r) < \text{BF}_{10;\kappa}(n, r)$.

Example 2 (*One-Sided Correlation Test for the US President Data, Continued*)**.** As shown in Fig. 6, for the Stulp et al. (2013) data the one-sided Jeffreys's exact correlation Bayes factor equation (27) yields $\text{BF}_{+0;\kappa=1} = 11.87$, indicating that the observed data are 11.87 times more likely under $\mathcal{M}_+$ than under $\mathcal{M}_0$. Because almost all posterior mass obeys the order-restriction, $\text{BF}_{+0} \approx 2 \times \text{BF}_{10}$—its theoretical maximum. ◇

Using the same arguments as above, we can define the Bayes factor for a test between $\mathcal{M}_-$ and $\mathcal{M}_0$, which is in fact given by $\text{BF}_{-0;\kappa}(n, r) = \text{BF}_{+0;\kappa}(n, -r)$ due to the fact that B is an odd function of ρ. In effect, this implies that $\text{BF}_{+0;\kappa}(n, r) + \text{BF}_{-0;\kappa}(n, r) = 2 \times \text{BF}_{10;\kappa}(n, r)$, where the factor of two follows from symmetry of $\pi_1(\rho\,;\,\kappa)$ in the definition of $\pi_+(\rho\,;\,\kappa)$. Additional information on the coherence (Mulder, 2014) of the Bayes factor for order restrictions is available elsewhere in this special issue (e.g., Mulder, 2016).

4.8. Discussion on the correlation test

As mentioned earlier, the previous analysis cannot be found in Jeffreys (1961) as Jeffreys did not derive the functions A and B explicitly. In particular, Jeffreys (1961, Eqn. (8, 9), p. 291) suggested

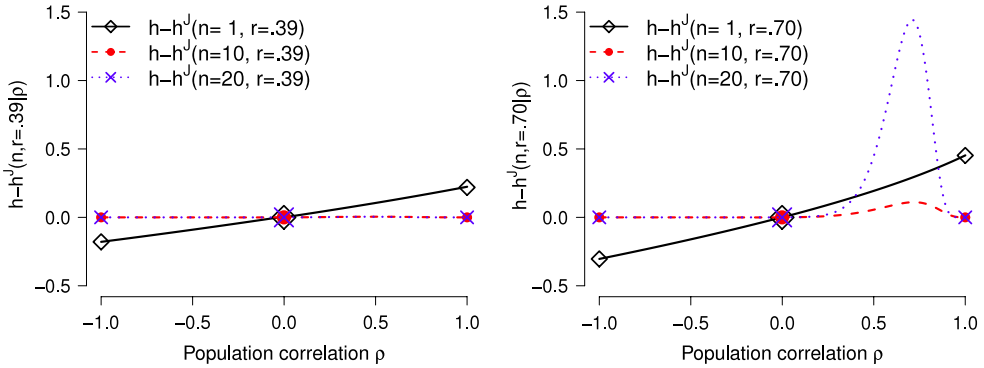

Fig. 7. Error of approximation between the exact function h and Jeffreys's approximation h^J. The left panel shows that for a modest sample correlation (i.e., $r = .39$, as in the example on the height of US presidents) Jeffreys's approximation is quite accurate; moreover, the error decreases as n grows, and the curve of $n = 10$ overlaps with that of $n = 20$. However, the right panel shows that for a sample correlation of $r = .70$ the error increases with n, but only for some values of ρ. Furthermore, note that Jeffreys's approximation h^J does not yield $h^J(n = 1, r) = 1$ for every possible r.

Table 2

A comparison of Jeffreys's exact Bayes factor (i.e., $\text{BF}_{10;\kappa=1}$) to Jeffreys's approximate integrated Bayes factor (i.e., $\text{BF}^{\text{J,I}}_{10}$) and to Jeffreys approximation of the approximate integrated Bayes factor (i.e., $\text{BF}^{\text{J}}_{10}$) reveals the high accuracy of the approximations, even for large values of r.

n	$\text{BF}_{10;\kappa=1}(n, .7)$	$\text{BF}^{\text{J,I}}_{10}(n, .7)$	$\text{BF}^{\text{J}}_{10}(n, .7)$	$\text{BF}_{10;\kappa=1}(n, .9)$	$\text{BF}^{\text{J,I}}_{10}(n, .9)$	$\text{BF}^{\text{J}}_{10}(n, .9)$
5	1.1	1.1	0.9	2.8	2.8	1.5
10	3.6	3.6	3.2	84.6	83.7	62.7
20	67.5	67.2	63.7	197,753.0	196,698.0	171,571.5

that the integral of the likelihood Eq. (17) with respect to the translation-invariant parameters $\pi_0(\theta_0)$ yields

$$h^J(n, r \mid \rho) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{(1-r\rho)^{\frac{2n-3}{2}}}, \tag{29}$$

which in fact approximates the true test-relevant likelihood function $h = A + B$ very well for modest values of $|r|$ (cf. Jeffreys, 1961, p. 175)—this is illustrated in Fig. 7 which plots the error $h - h^J$. Specifically, the left panel of Fig. 7 shows that when $r = .39$, as in the example on the height of US presidents, there is virtually no error when $n = 10$. The right panel of Fig. 7, however, shows that when $r = .70$, the error increases with n, but only for values of ρ from about .30 to about .95. From Jeffreys's approximation h^J one can define Jeffreys's integrated Bayes factor (Boekel et al., 2015; Wagenmakers, Verhagen, & Ly, in press):

$$\begin{aligned}\text{BF}^{\text{J,I}}_{10}(n, r) &= \frac{1}{2}\int_{-1}^{1} h_J(n, r \mid \rho)\mathrm{d}\rho \\ &= \frac{\sqrt{\pi}}{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n+2}{2}\right)}\,{}_2F_1\left(\tfrac{2n-3}{4}, \tfrac{2n-1}{4}\,;\, \tfrac{n+2}{2}\,;\, r^2\right).\end{aligned} \tag{30}$$

Jeffreys (1961, p. 175) noticed the resulting hypergeometric function, but as these functions were hard to compute, Jeffreys went on to derive a practical approximation for the users of his Bayes factor. The final Bayes factor that Jeffreys recommended for the comparison $\mathcal{M}_1$ versus $\mathcal{M}_0$ is therefore an approximation of an approximation and given by

$$\text{BF}^{\text{J}}_{10}(n, r) = \sqrt{\frac{\pi}{2n-3}}(1-r^2)^{\frac{4-n}{2}}. \tag{31}$$

For the US presidents data from Example 2 all three Bayes factors yield virtually the same evidence (i.e., $\text{BF}_{10;\kappa=1}(n = 46, r = .39) = 6.331$, $\text{BF}^{\text{J,I}}_{10}(n = 46, r = .39) = 6.329$, and $\text{BF}^{\text{J}}_{10}(n = 46, r = .39) = 6.379$). Table 2 shows that the three Bayes factors generally produce similar outcomes, even for large values of r (cf. Robert et al., 2009). Jeffreys's approximation of an approximation turns out to be remarkably accurate, especially because there is rarely the need to determine the Bayes factor exactly. Jeffreys (1961, p. 432) remarks:

> In most of our problems we have asymptotic approximations to K [i.e., BF_{01}] when the number of observations is large. We do not need K with much accuracy. Its importance is that if $K > 1$ the null hypothesis is supported by the evidence; if K is much less than 1 the null hypothesis may be rejected. But K is not a physical magnitude. Its function is to grade the decisiveness of the evidence. It makes little difference to the null hypothesis whether the odds are 10 to 1 or 100 to 1 against it, and in practice no difference at all whether they are 10^4 or 10^{10} to 1 against it. In any case whatever alternative is most strongly supported will be set up as the hypothesis for use until further notice.

Hence, the main advantage of having obtained the exact Bayes factor based on the true test-relevant likelihood function h may be that it justifies Jeffreys's approximation $\text{BF}^{\text{J}}_{10}(n, r)$. The true function h also provides insight in the one-sided version of Jeffreys's test, and it provides a clearer narrative regarding Jeffreys's motivation in model selection and hypothesis testing in general. Moreover, it allows us to show that Jeffreys's exact Bayes factor is model selection consistent.

4.8.1. Model selection consistency

To show that Jeffreys's correlation Bayes factor is model selection consistent, we use the sampling distribution of the maximum likelihood estimate (MLE). As r is the MLE we know that it is asymptotically normal with mean ρ and variance $\frac{1}{n(1-\rho^2)^2}$, where ρ is the true value. In particular, when the data are generated under $\mathcal{M}_0$, thus, $\rho = 0$, we know that $r \sim \mathcal{N}\left(0, \frac{1}{n}\right)$ when n is large. In order to show that the support for a true $\mathcal{M}_0$ grows without bound as the number of data points n increases, the Bayes factor $\text{BF}_{10;\kappa}(n, r)$ needs to approach zero as n increases.

We exploit the smoothness of $\text{BF}_{10;\kappa}(n, r)$ by Taylor expanding it up to third order in r. By noting that the leading term of the Taylor expansion $\text{BF}_{10;\kappa}(n, 0)$ has a factor $\Gamma\left(\frac{(n-1)\kappa+2}{2\kappa}\right)/\Gamma\left(\frac{n\kappa+2}{2\kappa}\right)$ we conclude that it converges to zero as n grows. The proof that

the Bayes factor $BF_{10;\kappa}$ is also model selection consistent under $\mathcal{M}_1$ follows a similar approach by a Taylor approximation of second order and consequently concluding that $BF_{10;\kappa}(n, r)$ diverges to ∞ as n grows indefinitely.

5. Conclusion

We hope to have demonstrated that the Bayes factors proposed by Harold Jeffreys have a solid theoretical basis, and, moreover, that they can be used in empirical practice to answer one particularly pressing question: what is the degree to which the data support either the null hypothesis $\mathcal{M}_0$ or the alternative hypothesis $\mathcal{M}_1$? As stated by Jeffreys (1961, p. 302):

> "In induction there is no harm in being occasionally wrong; it is inevitable that we shall be. But there is harm in stating results in such a form that they do not represent the evidence available at the time when they are stated, or make it impossible for future workers to make the best use of that evidence".

It is not clear to us what inferential procedures other than the Bayes factor are able to represent evidence for $\mathcal{M}_0$ versus $\mathcal{M}_1$. After all, the Bayes factor follows directly from probability theory, and this ensures that is obeys fundamental principles of coherence and common sense (e.g., Wagenmakers, Lee, Rouder, & Morey, 2014).

It needs to be acknowledged that the Bayes factor has been subjected to numerous critiques. Here we discuss two. First, one may object that the test-relevant prior distribution for the parameter of interest has an overly large influence on the Bayes factor (Liu & Aitkin, 2008). In particular, uninformative, overly wide priors result in an undue preference for $\mathcal{M}_0$, a fact that Jeffreys recognized at an early stage. The most principled response to this critique is that the selection of appropriate priors is an inherent part of model specification. Indeed, the prior offers an opportunity for the implementation of substantively different model (Vanpaemel, 2010).

In this manuscript, we showcased this ability when we adjusted the prior to implement a directional, one-sided alternative hypothesis. In general, the fact that different priors result in different Bayes factors should not come as a surprise. As stated by Jeffreys (1961, p. x):

> "The most beneficial result that I can hope for as a consequence of this work is that more attention will be paid to the precise statement of the alternatives involved in the questions asked. It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude".

The second critique is that in practice, all models are wrong. At first glance this appears not to be a problem, as the Bayes factor quantifies the support for $\mathcal{M}_0$ versus $\mathcal{M}_1$, regardless of whether these models are correct. However, it is important to realize that the Bayes factor is a relative measure of support. A Bayes factor of $BF_{10} = 100{,}000$ indicates that $\mathcal{M}_1$ receives much more support from the data than does $\mathcal{M}_0$, but this does not mean that $\mathcal{M}_1$ is any good in an absolute sense (e.g., Andraszewicz et al., 2015; Anscombe, 1973). In addition, it has recently been suggested that when both models are misspecified, the Bayes factor may perform poorly in the sense that it is too slow to select the best model (van Erven, Grünwald, & de Rooij, 2012). However, the Bayes factor does have a predictive interpretation that does not depend on one of the model being true (Wagenmakers, Grünwald, & Steyvers, 2006); similarly, the model preferred by the Bayes factor will be closest (with respect to the Kullback–Leibler divergence) to the true data-generating model (Berger, 1985; Jeffreys, 1980). More work on this topic is desired and expected.

In mathematical psychology, the Bayes factor is a relatively popular method of model selection, as it automatically balances the tension between parsimony and goodness-of-fit, thereby safeguarding the researcher against overfitting the data and preferring models that are good at describing the obtained data, but poor at generalizing and prediction (Myung, Forster, & Browne, 2000; Myung & Pitt, 1997; Wagenmakers & Waldorp, 2006). Nevertheless, with the recent exception of the Bayes factor t-test, the Bayes factors proposed by Jeffreys (1961) have not received much attention, neither by statisticians nor mathematical psychologists. One of the reasons for this unfortunate fact is that Jeffreys notation is more accustomed to philosophers of logic (Geisser, 1980). In order to make Jeffreys's work somewhat more accessible, Appendix D provides a table with a modern-day translation of Jeffreys's notation. In addition, any scholar new to the work of Jeffreys is recommended to first read the extended modern summary by Robert et al. (2009).

We would like to stress that a Jeffreys Bayes factor is not a mere ratio of likelihood functions averaged with respect to a subjective elicited prior $\pi_i(\theta_i)$ obtained from a within-model perspective. Jeffreys's development of the Bayes factor resembles an experimental design for which one studies where the likelihood functions overlap, how they differ, and in what way the difference can be apparent from the data. These consideration then yield priors from which a Bayes factor needs to be computed. The computations are typically hard to perform and might not yield analytical results. These computational issues were a major obstacle for the Bayesian community, however, Jeffreys understood that analytical solutions are not always necessary for good inference; moreover, he was able to derive approximate Bayes factors, allowing his exposition of Bayesian inductive reasoning to transcend from a philosophical debate into practical tools for scientific scrutiny.

Modern-day statisticians and mathematical psychologists may lack Jeffreys's talent to develop default Bayes factors, but we are fortunate enough to live in a time in which computer-driven sampling methods known as Markov chain Monte Carlo (MCMC: e.g., Gamerman & Lopes, 2006; Geisser, 1996) are widely available. This removes the computational obstacles one needs to resolve after the priors are specified. These tools makes Jeffreys's method of testing more attainable than ever before.

Acknowledgments

This work was supported by the starting grant "Bayes or Bust" awarded by the European Research Council. Correspondence concerning this article may be addressed to Alexander Ly. We thank Joris Mulder, Christian Robert, and an anonymous reviewer for their constructive comments on an earlier draft.

Appendix A. The default Bayes factor hypothesis tests proposed by Jeffreys in ToP

See Table A.3

Appendix B. Hypergeometric functions

The hypergeometric function (Oberhettinger, 1972, section 15) with two upper parameters and one lower parameter generalizes the exponential function as follows (Gradshteyn & Ryzhik, 2007, p 1005):

$$ {}_2F_1\,(a, b\,;\, c\,;\, z) = 1 + \frac{a \cdot b}{c \cdot 1} z + \frac{a(a+1)b(b+1)}{c(c+1) \cdot 1 \cdot 2} z^2 + \frac{a(a+1)(a+2)b(b+1)(b+2)}{c(c+1)(c+2) \cdot 1 \cdot 2 \cdot 3} z^3 + \cdots . \tag{B.1} $$

Table A.3
Default Bayes factor hypothesis tests proposed by Jeffreys (1961) in Chapter V of "Theory of Probability" (third edition).

Tests	Pages
Binomial rate	256–257
Simple contingency	259–265
Consistency of two Poisson parameters	267–268
Whether the true value in the normal law is zero, σ unknown	268–274
Whether a true value is zero, σ known	274
Whether two true values are equal, standard errors known	278–280
Whether two location parameters are the same, standard errors not supposed equal	280–281
Whether a standard error has a suggested value σ_0	281–283
Agreement of two estimated standard errors	283–285
Both the standard error and the location parameter	285–289
Comparison of a correlation coefficient with a suggested value	289–292
Comparison of correlations	293–295
The intraclass correlation coefficient	295–300
The normal law of error	314–319
Independence in rare events	319–322

Table D.4
Translation of the notation introduced by (Jeffreys, 1961, pp. 245–267). The treatment of α and β as new or old parameters differs from context to context in (Jeffreys, 1961).

Jeffreys's notation	Modern notation	Interpretation
q	$\mathcal{M}_0$	Null hypothesis or null model
q'	$\mathcal{M}_1$	Alternative hypothesis or alternative model
H		Background information (mnemonic: "history")
$P(q \mid H)$	$P(\mathcal{M}_0)$	Prior probability of the null model
$\int f(\alpha)\mathrm{d}\alpha$	$\int \pi(\theta)\mathrm{d}\theta$	Prior density on the parameter θ
$P(q'\mathrm{d}\alpha \mid H)$	$P(\mathcal{M}_1, \theta)$	Probability of the alternative model and its parameter
$P(q \mid aH)$	$\pi_0(\theta_0 \mid x)$	Posterior density on the parameter within $\mathcal{M}_0$
$P(q'\mathrm{d}\alpha \mid aH)$	$\pi_1(\theta_1 \mid x)$	Posterior density on the parameter within $\mathcal{M}_1$
K	$\mathrm{BF}_{01}(d)$	The Bayes factor in favor of the null over the alternative
α', β	$\theta_0 = \alpha, \theta_1 = \binom{\alpha'}{\beta}$	"Alternative" parameter $\theta_1 = \binom{\text{function of the old parameter}}{\text{new parameter}}$
$f(\beta, \alpha')$	$\pi_1(\eta \mid \theta_0)$	Prior of the new given the old prior within $\mathcal{M}_1$
$g_{\alpha\alpha}\mathrm{d}\alpha'^2 + g_{\beta,\beta}\mathrm{d}\beta^2$	$I(\vec{\theta})$	Fisher information matrix
$P(q, db \mid H) = f(b)db$	$\pi_0(\theta_0)$	Prior density of the common parameters within $\mathcal{M}_0$
$P(q'dbd\alpha \mid H) = f(b)dbd\alpha$	$\pi_1(\theta_1)$	Prior density of the parameters within $\mathcal{M}_1$
$P(\theta \mid q, b, H)$	$f(d \mid \theta_0, \mathcal{M}_0)$	The likelihood under $\mathcal{M}_0$
$P(\theta \mid q', b, \alpha, H)$	$f(d \mid \theta_0, \eta, \mathcal{M}_1)$	Likelihood under $\mathcal{M}_1$
$P(q\,db \mid \theta H)$	$\pi_0(\theta_0 \mid d)$	Posterior of the parameters within $\mathcal{M}_0$
$P(q'\,db\,d\alpha \mid \theta H)$	$\pi_1(\theta_1 \mid d)$	Posterior of the parameters within $\mathcal{M}_1$

Appendix C. The stretched beta density

By the change of variable formula, we obtain the stretched beta density of ρ on $(-1, 1)$ with parameters $\alpha, \beta > 0$

$$\frac{1}{2\mathcal{B}(\alpha, \beta)}\left(\frac{\rho+1}{2}\right)^{\alpha-1}\left(\frac{1-\rho}{2}\right)^{\beta-1}, \tag{C.1}$$

where $\mathcal{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function that generalizes $\binom{n}{k}$ to real numbers. By setting $\beta = \alpha$ this yields the symmetric beta density of ρ on $(-1, 1)$ with parameters $\alpha > 0$

$$\frac{2^{-2\alpha+1}}{\mathcal{B}(\alpha, \alpha)}(1-\rho^2)^{\alpha-1}. \tag{C.2}$$

The reparametrization we used in text is given by simply substituting $\alpha = 1/\kappa$ allowing us to interpret κ as a scale parameter.

Appendix D. Translation of Jeffreys's notation in ToP

See Table D.4

References

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R. P. P. P., Verhagen, A. J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*, 521–543.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534.

Bayarri, M., Berger, J., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of statistics*, *40*(3), 1550–1577.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Verlag.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, Vol. 1* (2nd ed.) (pp. 378–386). Hoboken, NJ: Wiley.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, *33*, 526–536.

Boekel, W., Wagenmakers, E.-J., Belay, L., Verhagen, A. J., Brown, S. D., & Forstmann, B. (2015). A purely confirmatory replication study of structural brain-behavior correlations. *Cortex*, *66*, 115–133.

Bolt, B. (1982). The constitution of the core: seismological evidence. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *306*(1492), 11–20.

Cook, A. (1990). Sir Harold Jeffreys. 2 April 1891–18 March 1989. *Biographical Memoirs of Fellows of the Royal Society*, *36*, 302–333.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.

Geisser, S. (1980). The contributions of Sir Harold Jeffreys to Bayesian inference. In A. Zellner, Kadane, & B. Joseph (Eds.), *Bayesian analysis in econometrics and statistics: essays in Honor of Harold Jeffreys* (pp. 13–20). Amsterdam: North-Holland.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.) (1996). *Markov chain Monte Carlo in practice*. Boca Raton (FL): Chapman & Hall/CRC.

Gino, F., & Wiltermuth, S. S. (2014). Evil genius? How dishonesty can lead to greater creativity. *Psychological Science*, *4*, 973–981.

Good, I. J. (1980). The contributions of Jeffreys to Bayesian statistics. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: essays in Honor of Harold Jeffreys* (pp. 21–34). Amsterdam, The Netherlands: North-Holland Publishing Company.

Gradshteyn, I. S., & Ryzhik, I. M. (2007). *Table of integrals, series, and products* (7 ed.). Academic Press.
Huzurbazar, V. S. (1991). Sir Harold Jeffreys: Recollections of a student. *Chance*, *4*(2), 18–21.
Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge: Cambridge University Press.
Jeffreys, H. (1924). *The earth, its origin, history and physical constitution*. Cambridge University Press.
Jeffreys, H. (1931). *Scientific inference*. Cambridge University Press.
Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society*, *31*(2), 203–222.
Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453–461.
Jeffreys, H. (1948). *Theory of probability* (2 ed.). Oxford, UK: Oxford University Press.
Jeffreys, H. (1955). The present position in probability theory. *The British Journal for the Philosophy of Science*, *5*, 275–289.
Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.
Jeffreys, H. (1973). *Scientific inference* (3 ed.). Cambridge, UK: Cambridge University Press.
Jeffreys, H. (1980). Some general points in probability theory. In A. Zellner, Kadane, & B. Joseph (Eds.), *Bayesian analysis in econometrics and statistics: essays in Honor of Harold Jeffreys* (pp. 451–453). Amsterdam: North-Holland.
Jeffreys, H., & Jeffreys, B. S. (1946). *Methods of mathematical physics*. Cambridge, UK: Cambridge University Press.
Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 19313–19317.
Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: a Bayesian approach. *Psychological Methods*, *10*(4), 477.
Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.
Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator. *Journal of the American Statistical Association*, *92*(438), 648–655.
Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481).
Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.
Lindley, D. V. (1980). Jeffreys's contribution to modern statistical thought. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: essays in Honor of Harold Jeffreys* (pp. 35–39). Amsterdam, The Netherlands: North-Holland Publishing Company.
Lindley, D. V. (1991). Sir Harold Jeffreys. *Chance*, *4*(2), 10–14. 21.
Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.
Ly, A., Marsman, M., Verhagen, A., Grasman, R., & Wagenmakers, E.-J. (2015). A tutorial on Fisher information. *Journal of Mathematical Psychology*, (submitted for publication).
Marin, J.-M., & Robert, C. P. (2010). On resolving the savage–dickey paradox. *Electronic Journal of Statistics*, *4*, 643–654.
Mednick, S. (1962). The associative basis of the creative process. *Psychological Review*, *69*(3), 220–232.
Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics & Probability Letters*, *92*, 121–124.
Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 153–171.
Mulder, J. (2016). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115.
Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906.
Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.
Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, *44*(1–2).
Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.
Oberhettinger, F. (1972). Hypergeometric functions. In M. Abramowitz, & I. A. Stegun (Eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (pp. 555–566). New York: Dover Publications.
O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics Vol. 2B: Bayesian inference* (2nd ed.). London: Arnold.
Pericchi, L. R., Liu, G., & Torres, D. (2008). Objective Bayes factors for informative hypotheses: Completing" the informative hypothesis and "splitting" the Bayes factor. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 131–154). New York: Springer Verlag.
Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite (Ed.), *The foundations of mathematics and other logical essays* (pp. 156–198). London: Kegan Paul.
Robert, C. P., Chopin, N., & Rousseau, J. (2009). Harold Jeffreys's theory of probability revisited. *Statistical Science*, 141–172.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.
Senn, S. (2009). Comment. *Statistical Science*, *24*(2), 185–186.
Stulp, G., Buunk, A. P., Verhulst, S., & Pollet, T. V. (2013). Tall claims? Sense and nonsense about the importance of height of US presidents. *The Leadership Quarterly*, *24*(1), 159–171.
Swirles, B. (1991). Harold Jeffreys: Some reminiscences. *Chance*, *4*(2), 22–23, 26.
van Erven, T., Grünwald, P., & de Rooij, S. (2012). Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society B*, *74*, 361–417.
Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.
Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.
Wagenmakers, E.-J., Lee, M. D., Rouder, J. N., & Morey, R. D. (2014). Another statistical paradox, or why intervals cannot be used for model comparison. (submitted for publication).
Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.
Wagenmakers, E.-J., Verhagen, A.J., & Ly, A. How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, (in press).
Wagenmakers, E.-J., Verhagen, A.J., Ly, A., Matzke, D., Steingroever, H., Rouder, J.N., & Morey, R.D. The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld, & I. Waldman (Eds.), *Psychological science under scrutiny: recent challenges and proposed solutions*. John Wiley and Sons, (in press).
Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, *50*(2).
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.
Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review*, *16*(4), 752–760.
Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.
Zellner, A. (1980). Introduction. In A. Zellner (Ed.), *Bayesian analysis in econometrics and statistics: essays in honor of Harold Jeffreys* (pp. 1–10). Amsterdam, The Netherlands: North-Holland Publishing Company.
Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, *6*, 233–243.
Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Jose, M. Bernardo, Morris, H. DeGroot, Dennis, V. Lindley, & Adrian, F. Smith (Eds.), *Bayesian statistics: proceedings of the first international meeting held in Valencia, vol. 1* (pp. 585–603). Springer.